**PES University, Bangalore**

(Established under Karnataka Act No. 16 of 2013)

**UE19CS203 – STATISTICS FOR DATA SCIENCE**

**EVALUATION SCHEME**

**Assessment Policy: In Semester Assessment (ISA) Max: 60(will be scaled down)**

**Assessment Policy: End Semester Assessment (ESA) Max: 100 (will be scaled down)**

**Total =ISA + ESA = 100 Marks**

| Activity | Marks | Remarks |
|---|---|---|
| ISA1 | 60 | Scaled Down to 25 |
| ISA2 | 40 | Scaled Down to 15 |
| Assignment | 40 | Scaled Down to 10 |
| **Total Marks** | **50** | |

| Activity | Marks | Remarks |
|---|---|---|
| ESA | 100 | Scaled Down to 50 |
| **Total Marks** | **50** | |

**PES University, Bangalore**

(Established under Karnataka Act No. 16 of 2013)

**UE19CS203 – STATISTICS FOR DATA SCIENCE**

# PROJECT GUIDELINES

## 1. Dataset Selection ( 5 Marks)

You are encouraged to select your own dataset.

The requirements are,

1. Dataset must have at least 500 or more observations and between 10 to 15 variables.
2. The dataset's variables should include categorical variables, discrete numerical variables, and continuous numerical variables.
3. There should be at least 3 – 5% of missing values and NULL or NA.

## 2. Exploratory Data Analysis (7 Marks)

Once you choose the dataset, do the following steps.

1. Describe your dataset. Explain the meaning of the columns that is there in the dataset.
2. Data Cleaning – Handle the missing data for both categorical and numerical variables (by dropping and imputing).
   (Note: The missing values cannot be just ignored or deleted without examining)
3. Remove unwanted observations – Duplicate/irrelevant/repetitive.
4. Fix the typos and inconsistent capitalization.

## 3. Graph Visualization (7 Marks)

1. Visualize the dataset to exhibit meaningful insights from it.
2. Use any three graph visualization techniques.
3. Filter unwanted outliers.
   a) Numerical – Box plot / Histogram.
   b) Categorical – Bar chart.

4. **Normalization and Standardization (5 Marks)**

    1. Compute the mean and variance for each of the columns.
    2. Normalize all the numeric columns, to make mean 0 and variance 1
    3. Discuss why is normalization is needed? How does it affect dataset?
    4. Use graphs used to check whether the data is normal.

5. **Hypothesis Testing (4 Marks)**

    1. State the research hypothesis.
    2. Perform statistical tests.
    3. Freedom to make your own hypothesis based on the columns.
    4. Decide whether the null hypothesis is supported or rejected.

6. **Correlation (3 Marks)**

    1. Find the correlation between variables that are positively and negatively related.
    2. State inferences about it.

7. **Presentation (5 Marks)**

    1. Presentation slides to be prepared.
    2. Screen shot of a particular part of dataset and related graphs can be put in PPT along with insights.

8. **Report (5 Marks)**

    A report of 4 to 5 pages needs to be submitted**.**

**Note: A team size of 4(Minimum number of students=Max. number of students=4) should be selected from your respective section. A Google sheet will be shared by your respective faculty to collect details about project.**

**Deadline for Project Details: 25th September 2020**

**Deadline for Project Completion: End of October 2020**

**Presentation Dates: 1st November 2020 to 10th November 2020.**